Data security strategies for next generation data warehouses

Safeguard your most complex data platforms with confidence

IBM

Introduction

Data warehouses deliver analytics with speed and simplicity to propel you ahead of the competition in the era of big data

Organizations are increasingly relying on data warehouses to provide the business critical reporting and analysis needed to empower business decisions and drive results. The data warehouse is a central repository for integrating data from one or more disparate sources such as applications, databases and legacy systems, a task that has become increasingly challenging to do with 2.5 quintillion bytes[1] of data created each day.

Critical to any big data strategy, data warehouses store current, as well as historical data, to satisfy business queries of all kinds. In the era of big data, a wide variety of data warehousing and analytic solutions enable the best approach for each business situation. From traditional multidimensional and data mining to mashups and industry models, you need the right solution for gleaning insight from information. Big data represents massive business possibilities and competitive advantages for organizations that are able to harness information, analyze it, and then turn it into business value. However, as the number of data warehouses storing sensitive data increases, we've also seen a rise in the number of attacks on those data warehouses. Organizations are challenged with implementing data security strategies to protect data.

Ensure data is secure

Trusting your data

The data stored in the warehouse is uploaded from operational systems such as marketing, sales, CRM and ERP, and lands in the warehouse via an extract-transform-load (ETL) process. Data inside the warehouse is cleaned, transformed, cataloged and made available for use by managers and other business professionals for data mining, online analytical processing, market research, security intelligence, decision support, and much more.

Do you know if data has been compromised as it speeds across enterprise systems and gets viewed by different users? Chances are you don't. According to the Verizon Data Breach Investigations Report[2], 66 percent of organizations do not know they have been breached until months after the incident. (See Figure 1) In 2012, 44 millions records of data were compromised, and over half came from database servers.

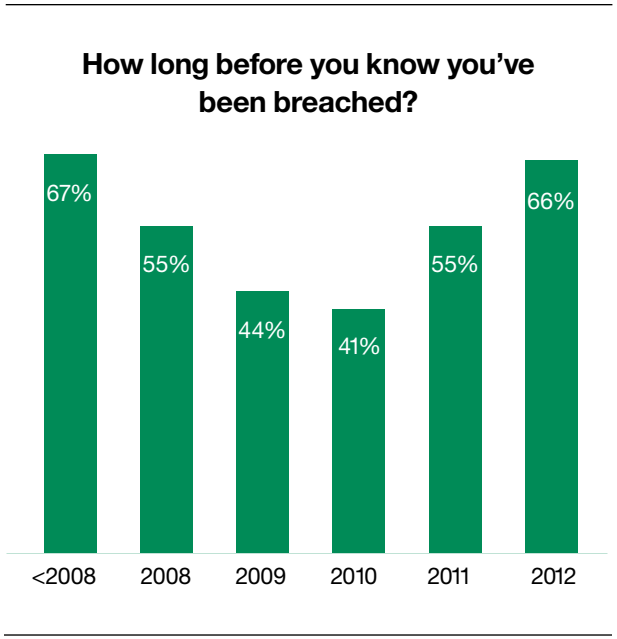

Figure 1: Percentage of data breaches taking months or more to discover

In an era of big data and high-volume, wide-variety and high-velocity data, organizations must be focused on improving accuracy, reliability and performance of the warehouse, because the market demands a real-time response. If the data is poor, unsecured or unavailable, even the most thorough analysis will not improve the bottom line. A recent survey (Business Case for Data Protection) indicated that preserving customer trust is one of the top motivations for protecting data. More and more customers are expecting companies to safeguard their information, but the impetuous doesn't stop here.

The costs of compliance and data breaches are devastating. Average compliance costs are estimated to be USD3.5 million[3], and the average cost of a security-related incident in the era of big data is estimated to be over USD40 million[4]. You can't afford to ignore data security and compliance as a top requirement.

Also startling is the number of breaches and intensity of breaches. The IBM X-Force 2012 Full Year Trend and Risk Report (March 2013) indicates that attacks are up 40 percent. (See Figure 2)

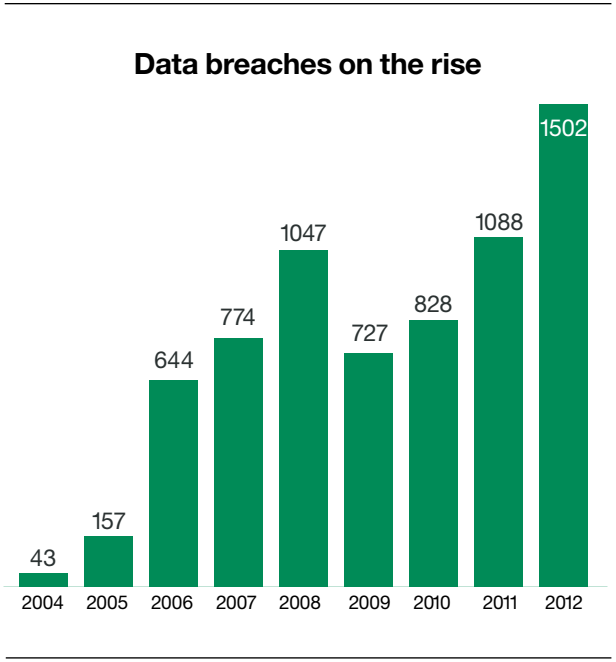

Figure 2: Data breach incidents over time

What data are attackers targeting?

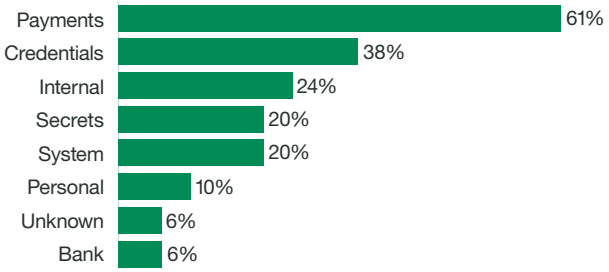

Figure 3: Break down of the data sought by attackers.

(*Source*: 2013 Data Breach Investigations Report, Verizon, April 2013)

Attackers are after payment data, system credentials, internal secrets and more. Unfortunately, only 15 percent of breaches had a complete and reliable count of compromised records, while the other 85 percent had no way to detect what data had been compromised[5] and more. (See Figure 3)

Many vendors offer a data warehouse platform such as Oracle, Teradata, Greenplum Database and IBM through IBM PureSystems™, and IBM provides data protection that can be seamlessly built into these platforms without impacting performance or slowing down analytics.

"Some organizations will be a target regardless of what they do, but most become a target because of what they do. If your organization is indeed a target of choice, understand as much as you can about what your opponent is likely to do and how far they are willing to go."

-2013 Verizon Data Beach Investigations Report

With the challenge of growing volume, velocity, and variety of data used today in all aspects of the business—using a multi-purpose system for all data workloads is often not the most cost effective or low risk approach, and definitely not the fastest to deploy. The new IBM PureData™ System is optimized exclusively for delivering data services to today's demanding applications. Like each of the IBM PureSystems, it offers built-in expertise, integration by design, and a simplified experience throughout its life cycle.

Built-in expertise
Codified data management best practices are provided for each workload. The PureData System delivers automated pattern-based deployment and management of highly reliable and scalable database services.

Integration by design
Hardware, storage and software capabilities are designed and optimized for specific high performance data workloads such as patented data filtering using programmable hardware (FPGAs) for ultrafast execution of analytic queries without the need for indices or complex partitioning schemes.

Simplified experience
The PureData System provides single part procurement with no assembly required (ready to load data in hours), open integration with 3rd party software, an integrated management console for the entire system, a single line of support, and an integrated system upgrades and maintenance.

The new PureData System comes in different models that have been designed, integrated and optimized to deliver data services to today's demanding applications with simplicity, speed & lower cost. IBM offers PureData for analytics, Hadoop, transactions and more with packages for industries.

Protecting the data warehouse

Analytics without security is an accident waiting to happen. What's at stake?

As you develop your data warehouse strategy, it is critical to include data security from the beginning. As you load data into the warehouse, could some of this data be sensitive? Do you have control over who or what is accessing this data? Do you know which data could fall under the scrutiny of an auditor? Are you or your competitors struggling with the effects of a breach or fine? What is your strategy for managing new risks and threats to the warehouse?

Now is the time to understand sensitive data and establish business-driven protection policies and controls to keep data safe without impacting performance. You will want to continuously monitor and audit data activity as well as de-indentify data as it moves into or out of the data warehouse, either through masking or encryption. In addition, policies need to keep up with the velocity of data—even one minute behind is too late. The longer you wait to apply a security policy, the greater your risk. With the average cost of a single data breach estimated to be USD5.4 million in 2013, delaying a security strategy is not an option.

The data warehouse consolidates and stores your most sensitive and critical business data. Perimeter security, firewalls, intrusion alerts and user access management may not detect when an advanced targeted attack has penetrated the organization, or when a super user misuses access privileges, makes unauthorized changes or mistakenly exposes data. The wide variety of data speeding through your enterprise and moving into and out of the data warehouse requires unprecedented levels of protection.

The bottom line: the increasing number of analytics systems storing sensitive data exponentially increases the risk of a breach.

What needs to be protected in the data warehouse environment?

An end-to-end security strategy is required for the data warehouse environment. A data warehouse environment consists of much more than just a database. The entire environment ranges from extracting data from operational systems, moving this data to the data warehouse, distributing the data to other analytic platforms, and finally, distributing it to the end business user.

In today's highly distributed, complex world, the environment spans multiple servers, applications and systems. When putting a security strategy in place, ask yourself "Who has a valid business need to know sensitive data?" If a user does not have a valid reason, then they should be denied access to the sensitive data. You might also ask yourself if sensitive data should even be entered into the data warehouse at all. In many cases, you can analyze trends at the aggregate level without compromising sensitive, detailed personal data that could break PHI/PII rules.

Here are some questions to ponder across three key areas of focus for data warehouse security:

Data inputs

- As data moves into the warehouse, how can you ensure integrity?
- Is a data classification policy in place? How is it applied to data entering the data warehouse?
- Does data even need to move to the data warehouse at this level?

Data outputs

- Is data exported from the warehouse to other applications, for example for reporting? If so, how is the data secured?
- How do you know that only authorized recipients are able to obtain the output?
- How do you know the right recipients receive the right information—and nothing more?

System security

- Have user access rights been determined and documented?
- Are they based on roles, such as through LDAP groups?
- Are administrator and super-user accounts carefully controlled and audited?
- Is the supporting database appropriately configured and hardened for maximum security?
- Is access to data restricted according to its' sensitivity?
- Does your monitoring strategy enable you to quickly detect and react when suspicious activity occurs?
- Are you auditing what you need to audit in order to fulfill the compliance mandates of your organization and/or industry? Can you quickly demonstrate the appropriate audit trails?
- After an attack happens, do you have the forensic tools to figure out what happened?
- How can you ensure security without impacting performance or real time analytics?

Today, a window of opportunity exists

Your clients demand accountability and visibility into how data is used and protected. You need to act fast, because data security isn't just a good idea; it is required by law.

"Disclosure laws mean that you can't keep quiet about a breach while you deal with the fallout. As well as trying to avoid being hacked in the first place, organizations need to be able to spot compromises quickly and minimize the amount of data lost."

—Verizon Data Breach Investigations Report.

IBM InfoSphere solutions scale to protect both traditional data management architectures and data warehouses against a complex threat landscape of internal and external threats. These threats can exploit vulnerabilities, resulting in the loss of confidentiality, integrity, or availability of a business asset.

Building security into big data environments can help satisfy a wide variety of mandates and integrate security with other IT management solutions such as leading authentication protocols, security information and event management (SIEM) solutions, ticketing systems, application servers, archival systems, and much more. The end goal is to improve security decision making based on prioritized, actionable insight without production impact.

A SIEM provides security intelligence for protecting assets and information from advanced threats.

IBM Security QRadar® SIEM consolidates log source event data from thousands of devices endpoints including data warehouses and applications distributed throughout a network. It performs immediate normalization and correlation activities on raw data to distinguish real threats from false positives. As an option, this software incorporates IBM Security X-Force® Threat Intelligence which supplies a list of potentially malicious IP addresses including malware hosts, spam sources and other threats. IBM Security QRadar SIEM can also correlate system vulnerabilities with event and network data, helping to prioritize security incidents.

Address the data security and compliance lifecycle with IBM InfoSphere solutions

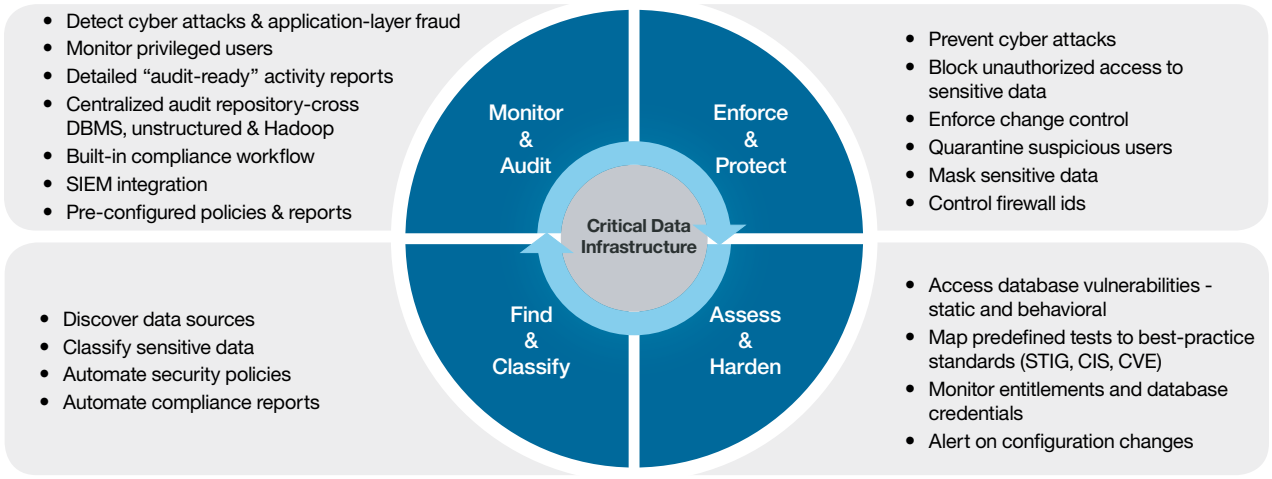

Figure 4. IBM InfoSphere Guardium addresses the security and compliance life cycle.

What does InfoSphere provide?

IBM InfoSphere solutions for data security provide a market-leading holistic protection approach for data warehouses, and deliver a 239 percent ROI in less than six months[6] without impacting performance. The portfolio consists of IBM InfoSphere Guardium® and IBM InfoSphere Optim™, and provides data access and change controls, real-time data monitoring and auditing, data protection and loss prevention, vulnerability management, and sensitive data discovery and classification to support compliance requirements and prevent breaches. InfoSphere protects data without impacting analytics or application performance.

The capabilities outlined below are available on all leading data warehouse platforms such as IBM PureData System, Teradata, Oracle and EMC Greenplum. Most leading data warehouse platforms come with built-in identity and access management systems. InfoSphere Guardium easily integrates with these existing security capabilities. (See Figure 4)

Find and classify

- Auto-discover and classify sensitive assets. Oversee access and entitlement management from a central location for all platforms.
- Set up and enforce policies to continually manage access to data warehouses.

Assess and harden

- Assess configurations and vulnerabilities and harden data warehouses.
- Assess data vulnerabilities through a variety of techniques; for example, check admin-level access privileges and verify that applications accessing data do not contain known vulnerabilities.
- Enable deeper insights to IT SIEM tools for more accurate and effective security intelligence.

Monitor and audit

- Monitor data stores to ensure that changes don't compromise security; if such changes do occur, alert the right people to take corrective action.
- Monitor activity into and out of data warehouse platforms.
- Vigilantly monitor user interaction with the data, detecting any unusual pattern of access, including privileged users and external access.
- Audit activity and report from a central location for all enterprise data warehouse platforms. To fulfill compliance requirements and ensure ability to perform forensic investigations, gather activity into nonrepudiable audit trails and appropriately formatted reports. Separation of duties is a key best practice here, since IT staff must not be able to tamper with reports about the systems they manage.

Enforce and protect

- Enforce data security policies with alerting and blocking for sensitive data access, privileged user actions, change control, application user activities, and security exceptions such as failed logins.
- Automate the entire compliance auditing process, including report distribution to oversight teams, sign-offs, and escalations with preconfigured reports relating to SOX, PCI DSS, and data privacy.
- Encrypt data at rest inside the data warehouse.
- Redact data for unstructured documents and forms: automatically recognize and remove sensitive content from unstructured data sources, such as scanned documents, PDFs, TIFFs, XML files and Microsoft® Word documents.
- Safeguard sensitive data while supporting information sharing for business use.
- Mask data as data moves into and out of the data warehouse. Mask data on demand to protect against abuse. Mask data wherever and whenever it appears across the enterprise. Some examples of where masking can be applied include data at rest or data in flight, relation data, flat files or data sets such as IBM IMS™ or VSAM, data in reports or documents, data moving into or out of the warehouse, and more.

IBM has pioneered a new approach to data masking, known as semantic masking, to keep pace with data privacy requirements in the new era of computing. This technology was developed in IBM's Zurich Research Lab to mask data in context based on rules to ensure accurate and consistent results for analytics. Semantic masking facilitates analytics and ensures sensitive data cannot be traced to an individual entity. The value of semantic masking is to retain the usefulness of the data while also adhering to compliance/regulation requirements.

Let's explore an example scenario

An international healthcare provider would like to conduct analysis on sensitive data from various sources across lifestyles, geographic regions and age groups. The international healthcare provider wants to know if financial status has anything to do with the propensity for a family member to get diabetes. They need a significant volume of patient data to be able to reach a conclusion across all subsidiaries worldwide.

Without semantic masking, international medical and insurance laws prohibit this type of sensitive data aggregation, because it can usually be tied to an individual entity. With a semantic masking solution, the international healthcare provider would be able to reach an outcome and provide analysis faster. Semantic masking algorithms ensure data is masked appropriately. For example, the algorithm would need determine if there are enough pancreatic cancers (Symptom Code 157) to summarize all 157.xx into 157. Semantically masked data will have the same symptoms and gender but the age, family income and ethnicity are intelligently masked to the proper range and to a valid set of data points. Therefore, it is impossible to identify a person or tie a person to income bracket, exact age, gender or ethnic background. Researchers achieve valid results while protecting privacy. An example of before and after semantic masking is shown in Figure 5.

Sementic masking

Original

SOURCE ID	NAME	HOME ADDRESS	SYMPTOM CODE	HOUSEHOLD INCOME	HOME PHONE	SEX	ETHNICITY	AGE
Europe	John Smith	5 Rue de la Paix Paris, France	157.0, 157.1, 157.2, 157.3, 185	75,000	01 58 71 12 34	M	Caucasian	43
Asia	Steve Jones	199 Huangpu Road Hongkou Shanghai, China	157.0, 185, 493.00, 493.02	44,000	021-6393 1234	M	Asian	21
North America	Angela Garcia	926 Central Reno, NV 89521	157.1, 185, 493.02	83,000	775.587.1578	F	Latino	47

Semantically Masked

SOURCE ID	NAME	HOME ADDRESS	SYMPTOM CODE	HOUSEHOLD INCOME	HOME PHONE	SEX	ETHNICITY	AGE
Europe	Jerry Jones	24 Boulevard Marlesherbes Paris, France	157, 185	79,500	01 55 27 12 34	M	Caucasian	41
Asia	Steve Smith	100 Century Avenue, Pudong, Shanghai, China	157, 185, 493	29,000	021-6322 5767	M	Asian	21
North America	Angela Hernandez	1142 Mainl Reno, NV 89521	157, 185, 493	68,500	775.773.1142	F	Latino	47

Figure 5. Semantic masking ensures sensitive data cannot be traced to an individual entity.

IBM InfoSphere Guardium

Benefits of continuous, real-time data security in your data warehouse

InfoSphere solutions for data security deliver deliver significant value to your data warehouse environment without impacting performance or big data analytics. (See Figure 6)

- **Prevent data breaches:** Avoid disclosure or leakage of sensitive data to mitigate the potential of a data breach, which cost about USD5.4 million per incident
- **Ensure data integrity:** Prevent unauthorized changes to data, data structures, configuration files and logs to ensure 100 percent visibility into data access patterns and trends
- **Reduce cost of compliance:** Automate and centralize controls and simplify the audit review process; one client deployed in less than 48 hours, and was able to conduct audits 20 percent faster
- **Protect privacy:** Prevent disclosure of sensitive information by masking or de-indentifying data in databases, applications, and reports on demand across the enterprise to save USD20 million in administrative overhead

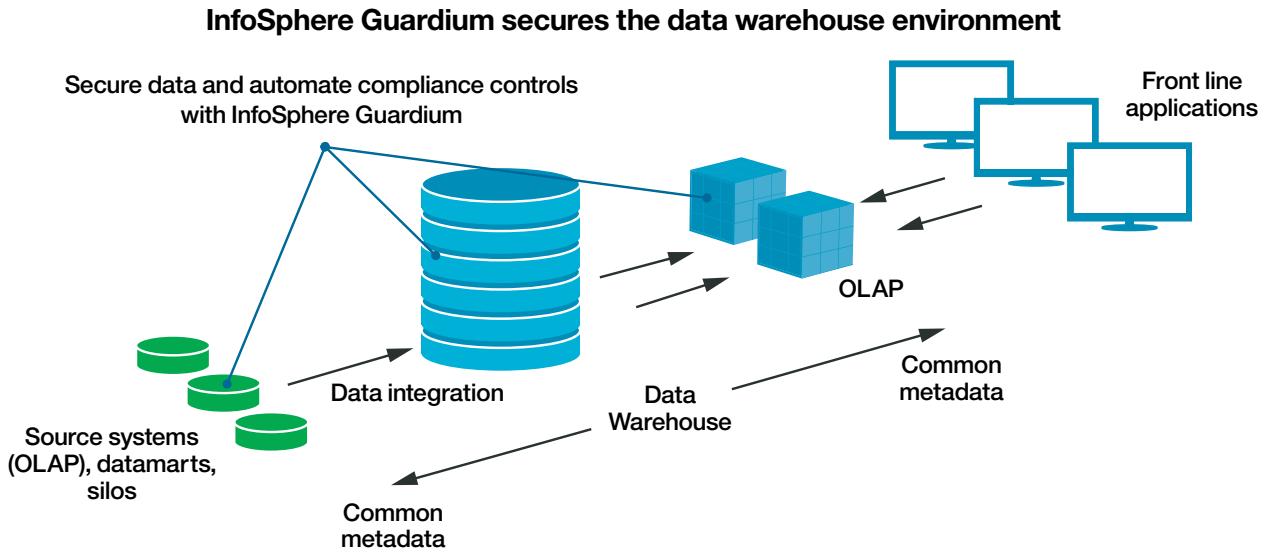

Figure 6. Data warehouse environment

Client Example: Protect data
A major global bank urgently needed to secure enterprise data and preserve data integrity across multiple business units, including the retail, corporate, investment and mortgage divisions. They needed to pass upcoming audits for multiple regulations including PCI-DSS (requirements 3, 6 and 10) and SOX. The data requiring protection included financial, HR, ERP, credit card, personally identifiable information (PII) and intellectual property. This data was housed in a diverse set of databases and warehouses including Oracle, SQL Server, Sybase, DB2 for LUW, DB2 for z/OS z, DB2 for i, Informix, MySQL and Teradata. Using InfoSphere Guardium Data Activity Monitor, the bank saved USD1.5 million per year in storage costs, since they no longer relied on native audit trails, and reduced security costs by USD20 million. No performance impact was noticed and InfoSphere Guardium Data Activity Monitor is now a standard part of the bank infrastructure, protecting against both internal and external threats. In addition, a culture shift has happened: there is now a greater awareness of the importance of data security.

About IBM InfoSphere Guardium

IBM InfoSphere delivers the confidence to act on big data

InfoSphere Guardium is part of the IBM InfoSphere integrated platform and the IBM Security Systems Framework. As the foundation of the IBM Big Data Platform, IBM InfoSphere provides market-leading functionality across all the capabilities of information integration and governance. InfoSphere creates confidence in big data by making it trusted and protected. InfoSphere is designed to handle big data: optimal scale and performance for massive volumes, agile and right-sized integration and governance for velocity, and for the variety to support many data types and big data systems. InfoSphere makes big data and analytics projects successful by delivering the confidence to act on insight.

InfoSphere capabilities include:

- **Metadata, business glossary and policy management:** Define metadata, business terminology, and governance policies with IBM InfoSphere Business Information Exchange
- **Data integration:** Handles all integration requirements, such as batch data transformation and movement (IBM InfoSphere Information Server), real-time replication (IBM InfoSphere Data Replication) and data federation (IBM InfoSphere Federation Server)
- **Data quality:** Parse, standardize, validate and match enterprise data with IBM InfoSphere Information Server for Data Quality
- **Master data management:** Act on a trusted view of your customers, products, suppliers, locations and accounts with InfoSphere MDM
- **Data lifecycle management:** Manage data lifecycle from test data creation through retirement and archiving with IBM InfoSphere Optim
- **Data security and privacy:** Continuously monitor data access, protect repositories from data breaches, and support compliance with IBM InfoSphere Guardium; ensure sensitive data is masked and protected with IBM InfoSphere Optim

For more information
To learn more about IBM InfoSphere Guardium solutions, contact your IBM sales representative or visit: **ibm.com**/guardium

Additional recommended reading

Hardening A Teradata Database: Best practices for access rights management

Protecting against database attacks and insider threats

8 Steps to Holistic Database Security

IBM Corporation
Software Group
Route 100
Somers, NY 10589

Produced in the United States of America
July 2013

1 ibm.com/software/data/bigdata

2 2013 Data Breach Investigations Report, Verizon, page 52, April 2013.

3 The True Cost of Compliance, Research Report, Ponemon Institute, January 2011.

4 The Big Data Imperative: Why Information Governance Must be Addressed Now, Research Brief, Aberdeen Group, December 2012.

5 2013 Data Breach Investigations Report, Verizon, April 2013

6 The Total Economic Impact of Guardium Database Security, Monitoring, and Auditing For A Global Consumer Products company. Forrester Consulting, January 2008.

Please Recycle